Project 1: Microarray Based Tumor Classification

Group: Saxophone
TA: Jing Zhang

Daniel Goldstein: Data Curator
Sooyoun Lee: Programmer
Yilin Yang: Analyst
Jason Rose: Biologist

INTRODUCTION

Colorectal cancer (CRC), also known as colon cancer, is often formed from benign polyps that become malignant over time. According to the American Cancer Society, CRC is the second leading cause of cancer-related death in the United States [1]. Prior studies sought to use microarrays to determine gene expression profiles (GEP) of CRC; however, due to the range of molecular subtypes of the disease, there is no single reliable signature for predicting prognosis [2,3]. GEP analysis using unsupervised hierarchical clustering and integrated genetic analysis has determined three distinct molecular subtypes of CRC [4]; therefore, we can further employ microarray technology to better understand the heterogeneity of CRC. In the study by Marisa *et al.*, unsupervised hierarchical clustering analysis revealed six molecular subtypes of CRC[5]. The aim of this project is to reproduce the analysis and results from Marisa *et al.*, comparing the C3 and C4 cancer subtypes.

DATA

The data used in Marisa et al. was collected by the Cartes d'Identité des Tumeurs (CIT) program in which 566 of 750 patients with stage I to IV CRC met the RNA quality requirements for GEP analysis. Data from the 566 patients were further subdivided into discovery (n = 443) and validation sample sets (n=123). An additional 903 samples were included in the validation set from several public datasets. For this project, we combined the discovery and validation sets for 134 samples. Expression profiling data was obtained through the GEO (www.ncbi.nlm.nih.gov/geo/) accession number GSE39582. The dataset met certain criteria for selection, including gene expression profile (GEP) data using the Affymetrix U133 Plus 2 chip with raw CEL files, as well as DNA alteration data for the following DNA mutations: microsatellite instability (MSI), CpG island methylator phenotype (CIMP), chromosomal instability (CIN), BRAF, and KRAS mutations. Since the subtypes identified in Marisa *et al.* were named according to their biological characteristics, subtypes C3 and C4 were named KRASm (*KRAS mutant*) and CSC (cancer stem cell), respectively. The CEL files were analyzed from our project directory on SCC (/projectnb/bf528/users/saxophone/project_1/samples). Symbolic links were created for 133 CEL files to limit redundancy among all groups and an additional file (GSM971958) was downloaded from the GEO database.

METHODS

Reading in Files

The following packages (affy, affyPLM, sva, AnnotationDbi, hgu133plus.db, ggplot2) were engaged for data pre-processing, quality control, and analysis in R version 4.0.3 and BiocManager 3.12. CEL files were read and stored with the ReadyAffy function from the affy package, and the AffyBatch was converted into a PLMset by fitting a specified robust linear model to the probe level data with the fitPLM function.

Normalization

In order to normalize the data, first, the CEL files were compressed and read into R using the ReadAffy function. Then by using the rma function, the variation between the arrays was corrected.

Quality Assurance

By using the Bioconductor package affyPLM the median of Relative Log Expression (RLE) and Normalized Unscaled Standard Error (NUSE) scores were summarized and computed. The RLE enabled us to assess whether the procedure aimed at removing unwanted variation and estimates of expression into the log scale for each gene and array [6]. Median values were then computed across the arrays for every gene. Most genes did not change in expression across samples, so the ideal median RLE values should be near 0. The NUSE normalized standard error estimates from the probe level model (PLM) and an ideal median standard error were close to 1 [7].

Batch Effects Correction

The ComBat package in R was used to adjust for batch effects [8]. The csv file named proj_metadata.csv, located in /project/bf528/project_1/doc/, was read. From this file, specific columns were used such as ‘normalizedcombatmod’ and ‘normalizedcombatbatch’. These two specific columns were extracted and used as an input variable to the ComBat function and the results were shown as the csv file.

Principal Component Analysis

The Principal Component Analysis (PCA) was performed to determine and visualize the variation present in a dataset with many variables [9]. The data was transposed to be able to scale and centered within each gene and the prcomp function was used to perform a principal components analysis. By using the ggplot function of the gglpot2 package, the PC1 vs PC2 plot was created [10]. The outliers were explained with the ggplot function and the boxplot for both PCA1 and PCA2 was created.

Noise Filtering and Dimensionality Reduction

In order to process and analyze the data effectively, noise filtering and dimensionality reduction was needed first. To achieve this, the dimensionality of the probes was reduced using three different types of filters. First, reading the expression data with read.csv returned a total of 54675 probes. Implementing the first filter returns the number of genes expressed in at least 20% of samples. This meant that for each gene, at least 20% of the gene-expression values must be > log2(15), so the original data was taken and after summing the number of rows that are greater than log2(15), the result returned 39661 probes.

The second filter involves a chi-square test, where the variance of each individual gene was calculated and computed the median of that variance. As for the test statistics, degrees of freedom was initially calculated - the number of columns in the original data minus 1, and turned out to be 133. Then, the median of variance of each row was taken to calculate the test statistic for each row to compare to the upper limit, which was calculated using chi_upper <- qchisq((1 - 0.99)/2, df, lower.tail = FALSE). After this level of filtering, there are 15508 remaining probes named as chi_filter.

For the third filter, the standard deviation and the means of the previous filtered data was first computed. The coefficient of variation was obtained through the equation sd/mean, and this value was filtered to be greater than 0.186 so that only those probes that have a coefficient of variation greater than 0.186 were left. There remained 1531 genes in total after this third level of filtering.

Clustering Analysis and Subtype Discovery

After having preprocessed and filtered the data, hierarchical clustering and subtype discovery of the data was performed. Hierarchical clustering is a form of unsupervised learning, which is a type of machine learning algorithm used to draw inferences from unlabeled data[12]. The hclust() function was used along with the Euclidean distance method to build a distance matrix. Then, the dendrogram was cut so that the samples are divided into two clusters, one having 57 genes and the other having 77 genes. A heatmap was then generated to illustrate a graphical representation of data that uses a system of color-coding to represent different values. This was accomplished by matching the data vector to the C3 subtype and appended the red or blue colors to the heatmap accordingly.

Lastly, a Welch t-test was performed to identify those genes that are differentially expressed between the two clusters. A Welch t-test is used to test the hypothesis that the two populations have equal means, thus the alternative hypothesis would be that they do not have equal means. First, the data were divided into two clusters/vectors to be sampled and computed the test statistic and adjusted p-value (also known as q-value). The overall number of genes significantly expressed in each cluster was 1249, which are known as differentially expressed genes. A gene is declared differentially expressed if a difference or change observed in read counts or expression levels/index between two experimental conditions is statistically significant. Statistical distributions, such as the Welch t-test implemented, are used to approximate the pattern of differential gene expression.

RESULTS

In this project, we have performed a quality control analysis to examine the reliability of the microarray data. After completing the process of the batch effect correction and normalization, the median of RLE and NUSE was computed by using the PLM function. The median of RLE and NUSE was shown as a histogram (Figure 1 and Figure 2).

The median of the RLE histogram is a useful tool to visualize the unwanted variation in high-dimensional data. An ideal median RLE value should be around 0 because the majority of the genes that were not biologically affected and the number of regulated genes was equal to the number of the down-regulated genes. In Figure 1, it has shown that the RLE medians are very close to 0 with a mean of 0.00265.

The median of the NUSE histogram showed from the standard errors that the median standard error across the arrays was equal to 1 for each gene. The low-quality arrays had median standard errors that were greater than 1. In Figure 2, a histogram of median NUSE showed the median value for NUSE was around 1or slightly greater than 1, and skewed right which represented a good expression of the array.

PCA was performed to determine and visualize the variation present in a dataset with many variables. The results were shown in Figure 3, as the first principal component (PC1) on the x-axis and the second principal component (PC2) on the y-axis. By observing this plot, there were no significant clusterings or outliers shown between the PC1 and PC2. To look for outliers in detail, the boxplot for PCA1 and PCA2 was visualized in Figure 4. The first principal component was shown on the left with no outliers however, the second principal component was shown on the right which showed three outliers that are away from the mean. And these three outliers that are in our dataset refer to GSM972097_050805-04.CEL.gz, GSM972350_MFL_036b_U133_2.CEL.gz, and GSM972467_MFL_400b_U133_2.CEL.gz.

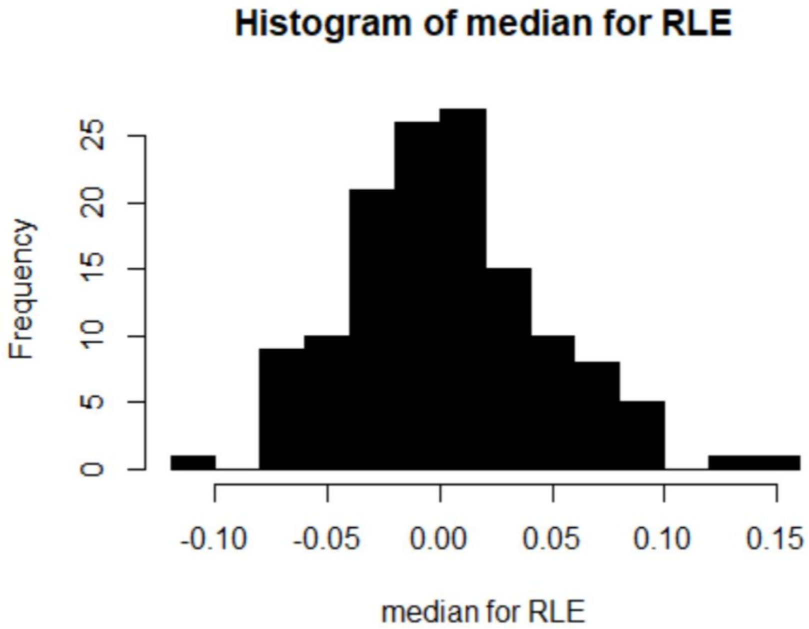

Figure 1 Histogram of the median for RLE. The RLE medians were measured across the patient samples. The majority of the patient samples were in between the RLE score of -0.05 to 0.10. The median which is larger than 0 represents that there is more gene expression in the samples that are upregulated.

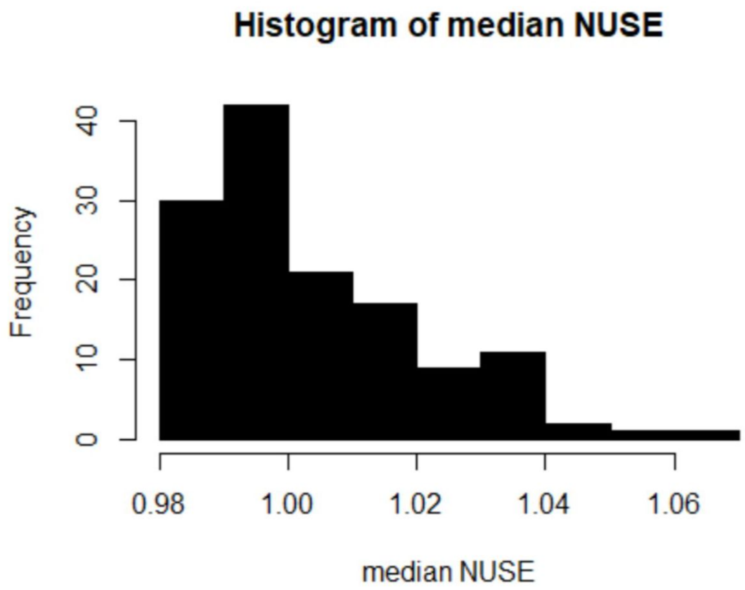

Figure 2 Histogram of the median NUSE. The NUSE medians were measured across the patient samples. This histogram enables us to identify lower quality arrays among the patient samples population. In this sample, the greater the NUSE score is, the lower the quality samples. But most of the sample median is in between the value of 0.98 to 1 which represents a good expression of the array.

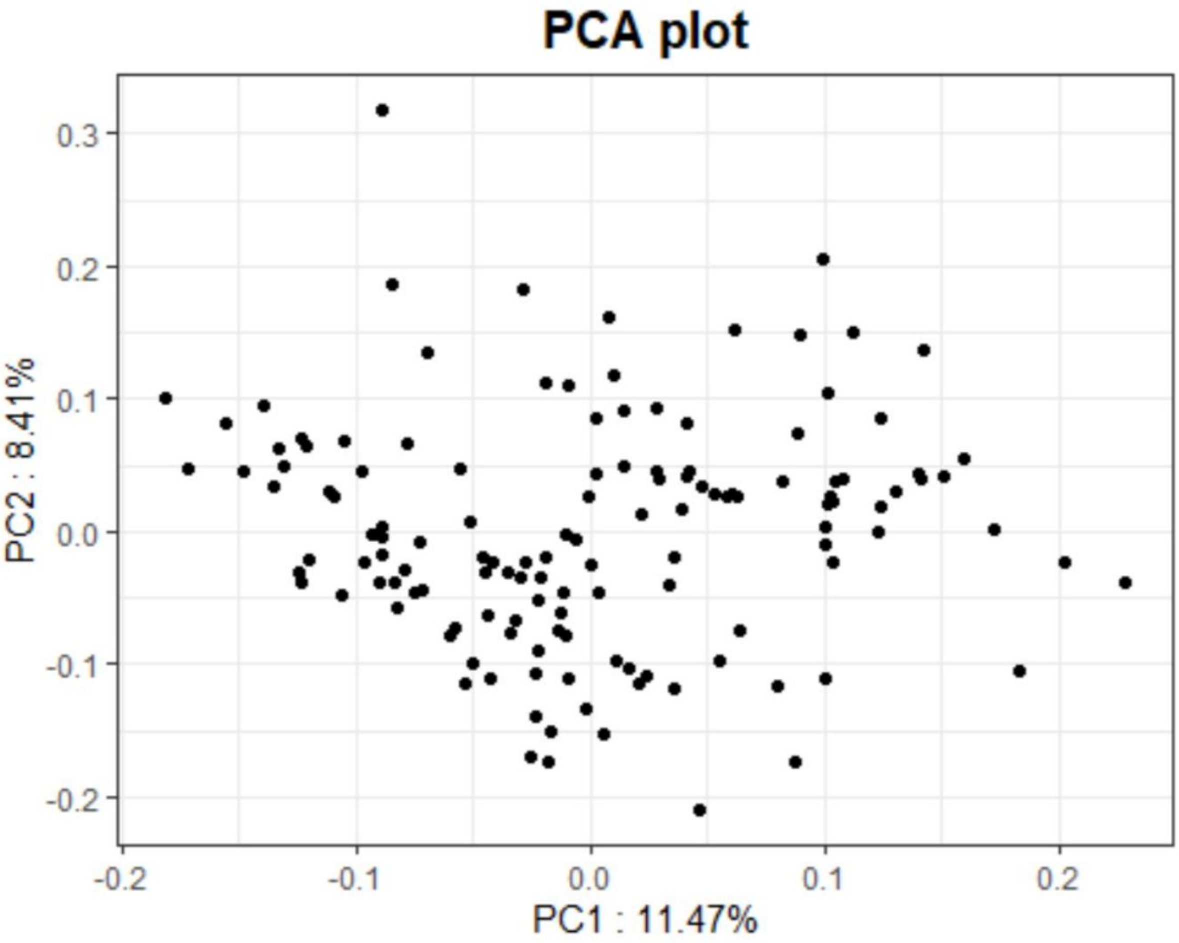

Figure 3 The PCA plot of patient samples. The first principal component (PC1) is represented on the x-axis and the second principal component (PC2) is represented on the y-axis. This PCA plot shows that there is no evidence that clustering exists between the PC1 and PC2. Also, the PC1 has 11.47% and the PC2 has 8.41% respectively. The 11.47% and the 8.41% of PC1 and PC2 represent the variance of the expression profile of 133 genes.

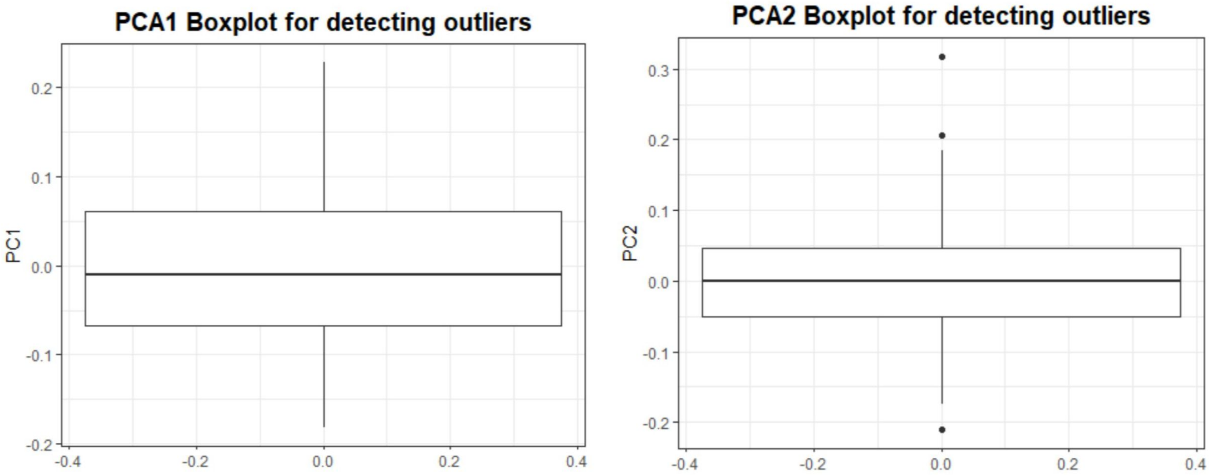

Figure 4 The Boxplot for detecting outlier in PCA1 and PCA2. This boxplot visualizes the outliers in PCA1 shown on the left, and PCA2 shown on the right. There were no outliers found in PCA1 which means all the patient samples are within the mean value. However, in PCA2, three outliers of patient samples were found which means these three outliers are not within the mean value.

We were able to obtain relatively close results to those found in the reference paper[1]. The original dataset obtained from data normalization contained 54675 probes in total, and the number of probes that passed through the expression filter were 39661 probes. After implementing the second filter (chi-square test), we were left with 15508 probes. Only 1531 genes remained as a result of the third coefficient of variation filter.

After analyzing the reduced data through hierarchical clustering, the data resulted in two clusters, which had 57 genes and 77 genes in each cluster respectively. A dendrogram, which is a diagram that shows the hierarchical relationship between objects was produced as a result of hierarchical clustering. A Welch t-test was then performed to give a data frame with probe set IDs, test statistic, p-values and q-values (also known as adjusted p-value). The resulting number of significant genes (p-val <0.5) was 1249.

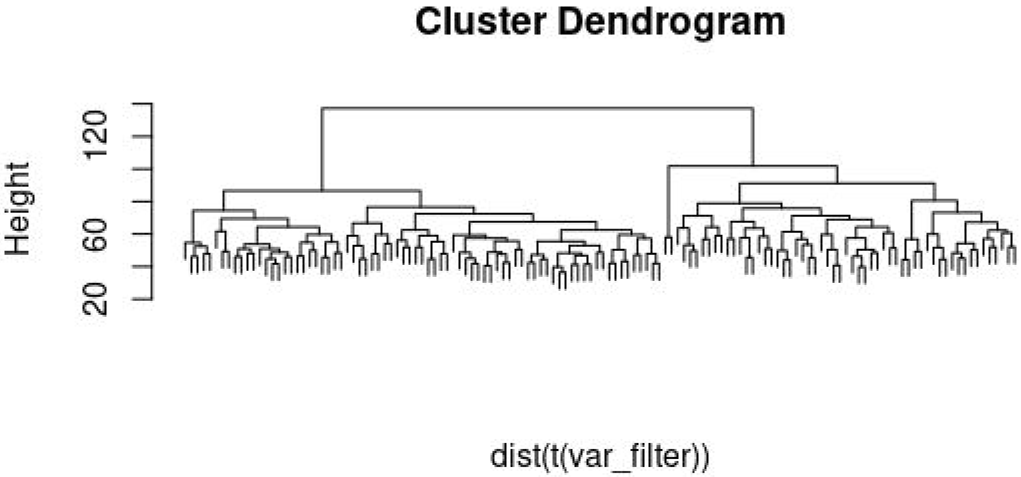

Figure 5 Dendrogram for hierarchical clustering. This Dendrogram shows the cluster result as obtained in 5.1. The matrix is divided into 2 clusters, with 57 genes in one cluster and 77 genes in the other. The R function hclust which calculates the dendrograms places the object (patients) labels at a constant distance below its clustering level.

The heatmap was created by matching the data vector to the C3 subtype and appended the red or blue colors to the heatmap accordingly. If the data matched the C3 subtype, blue was appended and vice versa. As for the list of differentially expressed genes that are most representative of each cluster, we chose the top 10 genes with the highest t-statistics are the most upregulated genes in the C4 subtype as compared to C3, and vice versa. As for the heatmap, correlation ranges from -1 to +1. Values closer to zero means there is no linear trend between the two variables. The close to 1 the correlation is the more positively correlated they are. Columns and rows of table are ordered so that similar genes and samples are next to each other.

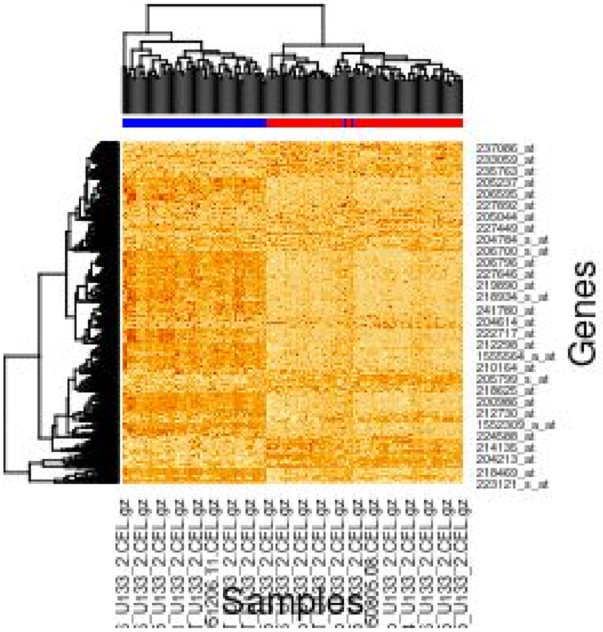

Figure 6 Heatmap as result of clustering. The red and blue colors in the top bar denote the C3 (75) and C4 (59) subtype samples respectively. All the samples in the cluster1 (77) are C4 subtypes. All the C3 (75) samples and only two remaining C4 samples are in cluster2.

A Welch t-test was then performed to give a data frame with probe set IDs, test statistic, p-values and q-values (also known as adjusted p-value). The resulting number of significant genes (p-val <0.5) was 1249. The heatmap was created by matching the data vector to the C3 subtype and appended the red or blue colors to the heatmap accordingly. If the data matched the C3 subtype, blue was appended and vice versa. As for the list of differentially expressed genes that are most representative of each cluster, we chose the top 10 genes with the highest t-statistics and the most upregulated genes in the C4 subtype as compared to C3, and vice versa.

The most differentially expressed genes that would match each cluster include 228233_at.t, 204720_s_at.t, 204818_at.t, 205184_at.t, 205525_at.t, and 206023_at.t, which are upregulated genes with positively high t-statistics as compared to 203404_at.t, 205529_s_at.t, and 218804_at.t which are downregulated and have negatively low t-statistic scores[13]. There were overall 1249 genes that are differentially expressed, with a q-value of less than 0.05. These genes may be significant in predicting and classifying the subtype of colon cancers.

DISCUSSION

Overall, the results obtained from this project were very similar to those in reference[1]. By using the ReadAffy function, the 134 CEL files were read, analyzed, and normalized. The AffyPLM function was used to be able to compute the samples into the two types of the histogram: the Relative Log Expression (RLE) and Normalized Unscaled Standard Error (NUSE). The data were corrected by doing the batch effect correction and the Principal

Component Analysis (PCA) was run to identify a higher tendency to be the outliers. We found that 97.76% (131 out of 134) of the samples were clustered correctly and three C3 subtype samples were classified together. For future implications, it would be interesting to find reasons why three outliers of the C3 subtype samples were classified and not being part of the C3 subtype.

We were able to utilize noise filtering to delete genes that do not express differentially across all the samples and obtain genes that could assist in distinguishing cancer subtypes. Combining these filtered data with unsupervised clustering methods enabled us to analyze the data more in-depth. The clustering results in our project demonstrate a high consistency with the results from reference [1]. The number of probes that passed all three filers, 1531, was very close to that in the supplementary materials. In addition, the number of genes in each clustered group (57 and 77) matched the paper's lists of genes used to assign subtypes as well. The top 10 deregulated genes include gene sets that represent epithelial-mesenchymal transition and metabolism of cytochrome P450, which were shown to be enriched in reference[1]. SFRP2, which is key to stem cell regulation and GAS1 growth arrest-specific 1, are down-regulated and were discussed in the paper as markers of aggressiveness of CC cells [1].

CONCLUSION

In this project, we have analyzed the microarray data of 134 colorectal cancer samples. The histogram for RLE and NUSE, PCA, and the boxplot was visualized. By creating the boxplot we have identified that three samples could be a possible outlier. We selected 1249 genes as the potential signature to classify or predict the C3 and C4 cancer subtypes.

REFERENCES

1. Siegel, R.L. *et al.* (2020). Colorectal cancer statistics, 2020. *CA Cancer J Clin.* 70(3): pp. 145-164. PMID: 32133645.
2. Wang, Y. *et al*. (2004). Gene expression profiles and molecular markers to predict recurrence of Dukes' B colon cancer. *J Clin Oncol* 22: 1564–1571.
3. Salazar, R. *et al.* (2011). Gene expression signature to improve prognosis prediction of stage II and III colorectal cancer. *J Clin Oncol* 29: 17–24.
4. Shen, L. *et al.* (2007). Integrated genetic and epigenetic analysis identifies three different subclasses of colon cancer. *Proc Natl Acad Sci USA* 104: 18654–18659.
5. Marisa *et al.* (2013). Gene Expression Classification of Colon Cancer into Molecular Subtypes: Characterization, Validation, and Prognostic Value. *PLoS Medicine* 10(5): pp. 1-13. PMID: 23700391.
6. Gandolfo, L. C. and Terence P. S. (2018). RLE Plots: Visualizing Unwanted Variation in High Dimensional Data. *PLOS ONE* 13(2). doi:10.1371/journal.pone.0191629.
7. Gentleman, R (2008). Bioinformatics and Computational Biology Solutions Using R and Bioconductor. Springer, pp. 45-47.

8. Leek, J. T. *et al.* (2012). The sva package for removing batch effects and other unwanted variation in high-throughput experiments. *Bioinformatics (Oxford, England)* 28(6): pp. 882-883. doi:10.1093/bioinformatics/bts034: 10-11.
9. Shah, I. (2018). Principal Component Analysis Utilizing R and SAS Softwares. *International Journal of Current Microbiology and Applied Sciences.* doi: 7. 10.20546/ijcmas.2018.705.441.
10. Holmes, S. and Huber, W. 3 High Quality Graphics in R. *Modern Statistics for Modern Biology*, Website: web.stanford.edu/class/bios221/book/Chap-Graphics.html.
11. Hierarchical clustering in R. (n.d.). Retrieved February 23, 2021, from https://www.datacamp.com/community/tutorials/hierarchical-clustering-R
12. What is hierarchical clustering? (2020, December 09). Retrieved February 23, 2021, from https://www.displayr.com/what-is-hierarchical-clustering/
13. Danielsson, F. *et al.* (2013). Majority of differentially expressed genes are down-regulated DURING malignant transformation in a four-stage model. *PNAS* 110(17): pp. 6853-6858. PMID: 23569271